キーワードの出現に基づく ブログコミュニティ抽出とオピニオンリーダーの発見

松永 拓† 平手 勇宇‡,†† 山名 早人‡,†††

†早稲田大学理工学部 〒169-8555 東京都新宿区大久保 3-4-1
††早稲田大学メディアネットワークセンター 〒169-8050 東京都新宿区戸塚 1-104
†††国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-1
E-mail: {taku,hirate,yamana}@yama.info.waseda.ac.jp

あらまし 近年，ブログは一般的に普及し，膨大な数が存在する．しかし，読み手が興味を持つ分野の情報を積極的に提供するようなブログはその中のごく少数である．こうしたごく少数の有用なブログに辿り着くためには，現状、キーワード検索や他サイトからのリンクなどを主とした偶然的な発見に頼らざるを得ない．本稿では，「あるコミュニティの話題を積極的に提供するようなブログ」を「オピニオンリーダー」と呼び，オピニオンリーダーを発見することを目的とする．従来、Web やブログを対象としたコミュニティ抽出では，ハイパーリンクのグラフ構造に基づいた手法が提案されている．しかし，ブログでは，身近な人へのリンクが多い，さまざまな話題への散発的なリンクが多い，という性質があり，ブログを対象とした場合，その精度を上げることが難しい．これに対して本稿では，ブログ上から，「興味の対象」を示すようなキーワードを抽出し，キーワードクラスタの抽出を行い，間接的にブログのコミュニティを抽出する．キーワードとしては「ニュース性のあるキーワード」、すなわち当該期間までほとんど出現せず，かつ,イベント性と重要性を兼ね備えるキーワードである．約 6.5 万件のブログを用いた実験の結果，興味別のコミュニティと，その興味の対象を中心的網羅的に扱うブログを得ることができることを確認した．

キーワード Web コミュニティ, ブログコミュニティ, オピニオンリーダー

Extracting Blog Communities and Opinion Leaders by using New Words Appearance

Taku MATSUNAGA† Yu HIRATE†‡ and Hayato YAMANA†,††

† Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan
‡ Media Network Center, Waseda University 1-104 Totsuka-cho, Shinjuku-ku, Tokyo, 169-8050, Japan
†† National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan

E-mail: {taku, hirate. Yamana}@yama.info.waseda.ac.jp

Abstract In recent years, blog has been known as a easy tool for putting out information, and many uses has begun to use blog. As a result, it is important to extract some kinds of knowledge from huge blog space. To extract some kinds of knowledge, conventional researches attempt to extract blog communities by applying web community extraction scheme by analyzing hyper links. This approach works well against blogs whose topics are constant. However, it does not work well against blogs whose topics are not constant. In this paper, to extract optimal communities against blogs of non-constant topics, we propose a novel scheme for extracting blog communities. Our scheme has following two new features. One is dividing one blog into number of topics, and extracts blog communities based on single topics instead of blogs. Due to this, one blog may belong to multiple communities. The other is employing “new words” based clustering as a topic clustering. We also propose a novel scheme for extracting opinion leaders from one blog communities. Our Evaluation using 65 thousands blogs shows that it extracts some small communities.

Keyword Web Community, Blog Community, Opinion Leader

1. はじめに

近年，ブログが普及し，多くの人々が書き手となった．総務省によれば，2005 年 3 月において，純ブログ利用者数は約 165 万人である[1]．膨大な数の情報源が得られるようになったが，全てを読むことは不可能であり，このような現状からブログを対象とした Web 解析の重要性が増してきている．

特に，ブログの持つパーソナリティから，トレンド分析のための研究が盛んに行われている．奥村ら[2]の blogWatcher では，Kleinberg[3] の手法を応用し，キーワードの注目度(burst 度)を得ることで，話題語を抽出する手法を提案している．また、同様にブログ全体から，話題語を抽出する試みとして Kizasi.jp[4]や，Technorati[5]が挙げられ，これらはすでに実用段階のサービスとして提供されている．しかし，読み手が興味のある話題を全般的に扱うブログを，大量のブログ群から，探し出すための手法は試みられてきていない．興味深いブログに辿り着くための方法は，現状ではキーワード検索や他サイトからのリンクなどを主とした偶然的な発見に頼っている．

本稿では，「あるコミュニティの話題を積極的に提供するようなブログ」を「オピニオンリーダー」と定義し，オピニオンリーダーを発見することを目的としている．そこで，ブログ全体をコミュニティ分析し，得られたコミュニティから，オピニオンリーダーを抽出する手法を提案する．オピニオンリーダーを抽出できれば，少数のオピニオンリーダーに注目するだけで，コミュニティ内の動向が得ることができ，コミュニティが注目するモノやイベント情報などの抽出に応用できると考えられる．

ブログデータ上から，重要なブロガーを発見しようとする研究としては，中島ら[6][7]の研究がある．中島らの研究では，ブログ記事同士のリンクをもとに得たブログトピックから，ブロガーが果たしている役割を特定することを目的としている．特に[7]の研究においては，Agitator と呼ばれる重要なブロガーを得る手法を提案している．Agitator は，Agitator のエントリーが加わることにより，ブログトピックに流れに何らかの強い影響を与えるブロガーである．これに対し，本稿での目的は，コミュニティの話題を網羅的に扱うブロガーを発見し，コミュニティの動向を得ることであるため，本稿でのオピニオンリーダーは議論に対しての影響力は考慮せず，コミュニティ情報の提供力のみで判別する．

従来，Web 上におけるコミュニティ分析においては，Web の特徴であるハイパーリンクのグラフ構造に注目した手法[8][10]が研究されてきた．しかし，これらの手法は従来的な Web サイトを対象としているため，性質の違いからブログに適用するには問題がある．谷口らの研究[11]では，従来の Web コミュニティ手法をブログへ適用することを試みている．この中で，興味深い問題点が 2 点報告されている．第 1 に，ブログホスティングサービス内部でのリンクが多く，ブログホスティングサービスのコミュニティが強く抽出されてしまう，という点である．結果として，「野球」のように長期にわたって同じテーマが記述されるようなコミュニティの抽出はうまくいくことがあるが，ブログホスティングサービスの繋がりが深く，話題の中心の発見はうまくいかなかった，と結論付けている．第 2 に，話題が多岐に渡るブログの扱いが困難である，という指摘をしている．例として，「たまに野球を語っている Blog は野球コミュニティよりも，ココログのコミュニティに入ってしまう」と述べている．多くのブログは一貫したテーマ性があるというよりは，個人的な出来事を記述しているため大きな問題である．解決策の方向性として，複数のコミュニティへの所属の許可やキーワードごとにコミュニティを抽出することを考察において挙げている．

そこで本稿では，コミュニティ抽出においては，上記の，ブログにおける Web コミュニティ抽出の問題点を解決する手法を提案する．具体的には，従来の Web コミュニティ抽出手法のようにブログ自体をクラスタリングせず，ブログからキーワードを抽出し，キーワードをクラスタリング対象とすることで，間接的にブログのクラスタリングを行う．キーワードは「興味の対象」であり，そのようなキーワードとして「ニュース性のあるキーワード」を抽出する．ここで，「ニュース性のあるキーワード」とは，まったくの新語である必要はないが，普段は出現しないキーワードであり，未知語も許容する，と定義する．キーワードをクラスタリングすることにより，ブログにおける Web コミュニティ抽出の問題点は次のように改善される．(1) 「興味の対象」を直接クラスタリングすることにより，コミュニティの「興

味」が明確に抽出できる．(2) クラスタリング対象がキーワードになるため，クラスタへの所属が適切に行われているかは，容易に判断が可能である．(3) ブログが複数のコミュニティへ所属することが可能になる．なお，ブログのキーワードクラスタへの所属と，キーワードクラスタへのオピニオンリーダーの所属の抽出は，ブログとキーワードクラスタとの類似性比較により行う．

本稿では次のような構成をとる．2 節では，関連研究について述べる．3 節では提案手法について述べる．4 節では実験を行った結果について述べる．5 節は本稿のまとめである．

2. 関連研究

2.1. Web コミュニティ

Web 空間から，同じ興味を共有するコミュニティの抽出が近年研究されている．(表 1)

表 1. 関連研究

発表年	研究	コミュニティの関係性	対象
1999	Kumar ら[8]	完全２部グラフ[10]	Web
2001	村田[9]	完全２部グラフ[10]	Web
2004	谷口ら[11]	リンク/トラックバックをエッジとして Betweenness クラスタリング[12]	ブログ
2006	内田ら[13]	トラックバックをエッジとし，Newman 法[16]でのクラスタリング	ブログの記事

Kumar ら[8]は Kleinberg[9]が提案した HITS アルゴリズムを用いて完全２部グラフとなっているコアをコミュニティとして抽出する手法を提案した．村田[9]は Kumar ら[8]の手法を発展させ，検索エンジンから backlink を得て，処理量を減らす手法を提案した．

ブログにおいては，谷口ら[11]は同様にリンク構造の関係性からコミュニティを抽出する手法を試みている．谷口ら[11]の研究では，ブログでは，ホスティングサービス内部でのつながりが強く，少なくともこのような特性を考慮しないと話題の中心が発見できないということが示された．また，話題が多岐に渡るブログの扱いへの問題提起がなされた．

ブログの記事単位においては，内田ら[13]が，ブログの記事をノード，トラックバックをエッジとしてクラスタリングを行うことにより，意味的にもまとまりのあるクラスタを得られることを示した．しかし，得られたクラスタは「新潟越中地震」「プロ野球参入問題」「年金制度問題」といったトピックごとのクラスタであり，オピニオンリーダーの抽出のためには粒が小さいと考えられる．

2.2. 重要なブロガーの発見

ブログデータ上から,重要なブロガーを発見しようとする研究としては，中島ら[7]の研究がある．中島らの研究では，ブログ記事同士のリンクをもとに得たブログトピックから，Agitator と呼ばれる重要なブロガーを得る試みをしている．Agitator の判定は，(1)あるブロガーのエントリーの被リンク数，(2) あるブロガーのエントリーによって，エントリー以前と以後のトピック内のエントリー数がどれくらい増加したか，(3) あるブロガーのエントリーと，エントリー以前のエントリー郡とエントリー以後のエントリー郡の類似性の変化，の３つの観点から議論をしている．中島らの目指す Agitator の取得とは，ブログトピックの流れに何らかの影響を与え，ブログトピックを変化させることのできるブロガーを得ることを目的としている． これに対し，本稿での目的は，コミュニティの話題を網羅的に扱うブロガーを発見し，コミュニティの動向を得ることである．そのため，本稿でのオピニオンリーダーは議論に対しての影響力は考慮せず，コミュニティ情報の提供力のみで判別する．

3. 提案手法

本節では，キーワード出現に基づくブログコミュニティ抽出とオピニオンリーダーの発見手法を提案する．提案手法は，ニュース性のある情報が得られる場（キーワードクラスタ）と,情報を提供している人（オピニオンリーダー）を抽出するものであり，最新情報を貪欲に得たいというユーザにとって，多くの機会を提供するという利点をもたらす．

以下 3.1 において既存手法の問題点，提案手法の有用性を示し，3.2 で提案手法の概要を説明する．そして 3.3 以降で提案手法の詳細について説明をする．

3.1 既存手法の問題点, 提案手法の有用性

Web コミュニティ抽出においては２章で述べたとおり、Web 空間から同じ興味を共有するコミュニティの抽出が研究されているが，ブログに適用する場合の問題点として，(1)ホスティングサービス内部でのリンクのつながりが強い (2) 話題の散発するブログのコミュニティへの所属の問題 が挙げられる．

提案手法では，コミュニティ対象を「ニュース性のあるキーワード」とすることで，リンク構造に頼らず

に話題を発見し，また，話題が散発するブログにおいてもコミュニティの評価を可能にした．

3.2 提案手法の概要

提案手法の流れは，図 1 の通りである．

(1) コミュニティ抽出の対象となる「ニュース性のあるキーワード」の抽出を行う（図 1 の(1)）．

(2) 抽出した「ニュース性のあるキーワード」を基に，キーワードクラスタを抽出する（図 1 の(2)）．

(3) (2)で抽出されたキーワードクラスタを基に，ブログをキーワードクラスタに紐付け，ブログコミュニティを抽出する（図 1 の(3)）

(4) 抽出されたキーワードクラスタを基にして，オピニオンリーダーを抽出する（図 1 の(4)）．

以後提案手法の詳細について述べる．処理(1)の「ニュース性のあるキーワード」の抽出方法を 3．3 で，処理(2)のキーワードクラスタの抽出手法について 3．4 で，処理(3)のブログコミュニティの抽出方法について 3．5 で，処理(4)のオピニオンリーダーの抽出を 3．6 でそれぞれ説明する．

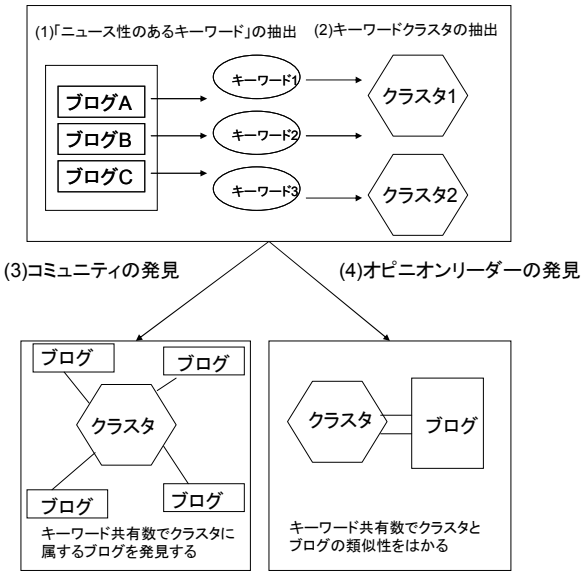

図 1．提案手法の流れ

3.3 ニュース性のあるキーワードの抽出

本節では，ブログデータから「ニュース性のあるキーワード」の抽出手法に関して説明する．本稿では，「ニュース性のあるキーワード」を以下の 3 つの条件を満たしたキーワードと定義する．ニュース性とは，「新規性」「イベント性」「重要性」の 3 点で構成される．

(1) 突発的に複数のブログで出現する（イベント性）

(2) (1)での出現以後は継続的に出現する（重要性）

(3) 普段は出現しない（新規性）

(1)は 24 時間とする．24 時間であれば，1 日のどの時点でイベントが発生しても，ブロガーが自分の執筆時間を決めていれば，補足可能であるからである．(2)に関しては，「最大出現期間」として，最初の出現時刻から最後に観測された出現時刻の差をパラメータ化し，実験において条件をいくつかのパターンとして用意して，もっとも良い結果を得られた条件を採用することとした．(3)は解析対象期間以前に十分な期間のデータをキーワードの未出現期間の判定のために用意することで判断できる．

提案手法では，このような「ニュース性のあるキーワード」の抽出のために，以下に示すような方法で抽出を行った．

(1) ブログのテキストデータを Mecab[14]を用いて，形態素解析を行い，名詞句を抽出し，「ニュース性のあるキーワード」候補リストに追加する．

(2) 「ニュース性のあるキーワード」は未知語の可能性があるために，適切に形態素に分解できない場合がある．そこで，名詞句が隣接する場合は，隣接する名詞句同士を結合した複合語も，「ニュース性のあるキーワード」候補リストに追加する．

(3) キーワードの正規化と記号等を含むキーワードの除外を行う．正規化を行うのは，同義語と思われるキーワードを統一するためである．Mecab[14]では，未知語推定によって，「♪」等が名詞と推定される場合がある．(2)の手順で示したとおり，「ニュース性のあるキーワード」の抽出手法では，未知語対応のために，隣接する名詞句同士を結合する．このため，同一名詞句（複合語も含む）も記号等が含まれることにより，違うキーワードとして認識してしまい，ブログの文章では，記号が文章中のアクセントとして使われる（「楽しい♪」など）場合が多いため，ノイズとなるからである．ルール一覧は[表 2]で示す．

(4) (1),(2),(3)で生成した「ニュース性のあるキーワード」候補リストから，「ニュース性のあるキーワード」の定義である条件(a)最小出現間隔 (b) 最大出現期間の判定を行う．条件(a)(b)を満たす単語を抽出し，「ニュース性のあるキーワード」とする．最小出現間隔は 24 時間以内とし，最大出現期間を「2 日~6 日以上で 1 日おき」の 5 パターンの実験を行う．

表 2 キーワードの正規化ルール一覧

ルール	詳細
数字・記号の除外	以下の数字・記号を含むキーワードを除外 0 1 2 3 4 5 6 7 8 9 < > [ ] ( ) { } : ; + - * % / * = ? ! . | & # @ _ ' " ~ ♪ ⇒ …. . ‘’ “” () 〔〕 [] {} 〈〉《》「」『』【】 ＋－±×÷＝≠ ＜＞≦≧∞∴♂♀° ′ ″ ℃ ＄ ¢ £ % ♯ & ＊ @ § ☆ ★ ○ ● ◎ ◇ ◆ □ ■ △ ▲ ▽ ▼ ※ 〒 → ← ↑ ↓ 〓 ∈ ∋ ⊆ ■ ⊇ ⊂ ⊃ ∪ ∩ ∧ ∨ ￢ ⇒ ⇔ ∀ ∃ ∠ ⊥ ■ ⌒ ∂ ∇ ≡ ≒ ≪ ≫ √ ∽ ∝ ∵ ∫ ∬ Å ■ ‰ ♯ ♭ ♪ † ‡ ¶ ◯ ― | ┌ ┐ ┘ └ ■ ├ ┬ ┤ ┴ ┼ ━ ┃ ┏ ┓ ┛ ┗ ┣ ┳ ┫ ■ ┻ ╋ ┠ ┯ ┨ ┷ ┿ ┝ ┰ ┥ ┸
接尾名詞の除外	末端に接尾語を含むキーワードを除外（Mecab の解析結果を利用） 「くん」「さん」「氏」など
日付を指す名詞の除外	「今日」「昨日」「明日」「明後日」 「今年」「本年」「来年」「昨年」 「去年」
英字の大文字小文字の統一	全て大文字に統一
カタカナ・ひらがなの統一	全てカタカナに統一
英字・数字・カタカナの全角・半角の統一	全て半角に統一

3.4 キーワードクラスタの抽出

3.1 で示した手法で得られた「ニュース性のあるキーワード」を Newman 法[10]を用いてクラスタリングを行う．Newman 法では，階層的クラスタリングの手法で，Modularity Q が高くなるクラスタを結合していき，Modularity Q が最大となるまで繰り返す．Modularity Q は，GN 法[11]で提案された指標で定義は次のとおりである．

e_{ij}をクラスタ i から j へのエッジの割合とし，$a_i = \sum_j (e_{ij})$としたとき，

$$Q = \sum_i (e_{ii} - a_i^2)$$

である．

Modularity Q は，1 に近づく程，良いクラスタが行われているといえ，0 では，ランダムな接続と同じである．

初期状態においては，抽出された「ニュース性のあるキーワード」が，抽出の際に関連付けられたブログのうち，共通のブログの数だけ，外部リンクを設けるように構成する．なお，この状態は，ブログが「ニュース性のあるキーワード」によってクラスタリングされているとも見ることができ，同じ「ニュース性のあるキーワード」に属するブログ同士で内部リンクを張り合うこともできるが，そのような構成の場合，初期状態から Modularity Q の値が高くなってしまため，内部リンクは設けないようにした．

3.5 所属の判定

ブログにおけるキーワードクラスタ所属の判定と，キーワードクラスタにおけるオピニオンリーダーの判定は，それぞれのブログが持つ「ニュース性のあるキーワード」の数を a，キーワードクラスタが持つ「ニュース性のあるキーワード」の数を b，ブログとキーワードクラスタで共通する「ニュース性のあるキーワード」の数を c，としたとき以下の条件で算出することができる．

(1)あるブログにおけるキーワードクラスタ所属の判定

対象ブログにおいて，c / a を，当該ブログのキーワードクラスタへの所属度とする．これは，当該ブログが，複数のキーワードクラスタに属する場合，どのキーワードクラスタに属するかを決定するための判定基準となる．

(2) キーワードクラスタにおけるオピニオンリーダーの判定

対象キーワードクラスタに属するブログの中で，c / b が最大となるブログを，当該キーワードクラスタのオピニオンリーダーとする．これは，キーワードクラスタを形成する「ニュース性のあるキーワード」リストを，最も多く含んでいるブログがオピニオンリーダーであることを意味する．

4. 実験

4.1 ブログデータの収集

Google Blogsearch[17]では，検索結果の RSS フォーマットを提供している．これを用いるとある検索語を含む最新のブログを得ることができる．Google Blogsearch を用いて，RSS フィードを収集し，RSS フィードからタイトル，発行時間，デスクリプションを含むブログデータを収集した．なお，検索語は検索結果が特定ジャンルに偏らなければ問題ないと考えられるので，今回は「あ」および「で」の 2 つの日本語文章に頻出するキーワードを用いた．これにより，2006/12/15～2007/1/15 の期間の記事データ，65,940 件を対象とした．

4.2 .キーワードクラスタ抽出

抽出した27,821個の「ニュース性のあるキーワード」に対して，Newman 法によるクラスタリングを行った．キーワードの抽出条件と Modularity Q 値の関係を示す[表 3]．最大出現期間が 4 日以上のとき Modularity Q 値は最大値をとり，その値は **0.5331** となった．

表 3．最大出現期間と Q 値

最大出現期間（日）	2	3	4	5	6
Q	0.5318	0.5331	0.5331	0.5307	0.5310

Modularity Q が最大の値を取るとき，すなわち最大出現期間が 4 日の時，キーワードクラスタは 7 個抽出された．7 個のうち，2 個はキーワードが 10 個未満の小さなコミュニティであった．[図 2]に大きな 5 つのコミュニティに推定される特徴を挙げる．また，キーワードの抜粋を[表 4]に挙げる．

図 2．全体コミュニティ（括弧内はキーワード数）
Q=0.5331

表 4．各コミュニティのキーワードの抜粋

女性コミュニティ 少年倶楽部プレミアム, NEWS ツアー, ミンテレ, ジャニカン, コン終了, クラクリスマス, 石川コン, 冬休ミドラマスペシャル, ジャニーズカウントダウンコンサート, クラクリスマス SP, カウコン以来, AHAPPYNEWSYEAR, NEWS コンサート, オリコン授賞, エイトドラマ, チャン出演, 全部録画, 部参戦, コン映像, 城ホ, 広島コン, テゴマス出演, NEWS 担, 胸宇宙, SHOCK 初日, 翼クリスマス, コン以外, 君出演, 翼クリスマスコンサート, カウコンレポ, コンチケ, 胸宇宙 ENDLICHERI, HARMONYOFWINTER, 元旦コン, ダブルアンコール, 淳太, MC レポ, FUTARI,
男性コミュニティ マイネルボウノット, 石橋守騎手, ダノンデインヒル, 着固定, 年京成, 山本茜騎手, ドリパス, キングスゾーン, 武幸四郎騎手, 番人気スリー, フサイチエアデール, マイネルポライト, オーゴンサンデー, メジロトンキニーズ牝, 頭サラ, 番人気ダイワ, センギョウシュフ, テンビー, クリスマス C, 勝浦正樹騎手, ゼンノエルシド, オースミヘネシー, 年 JRA, アクロスザヘイブン, 馬予想, スプリンターズ S, 長谷川浩, 安藤勝己騎手騎乗, 有馬記念優勝, 年連続年度代表, スカーレットブーケ, レイナワルツ, 師試験, 番人気スリーアベニュー, ディープインパクト武豊
オタク系 ハルヒ関連, 修正 DVD, 憂鬱 DVD, プーリップ, スペース長門, ギルティギアイグゼクス, アクセントコア, 姉チャンバラ, 姉チャンバラ VORTEX, 探偵コナン追憶, FFXII, プレサイトオープン, 七尾奈留, ローゼン麻生, 牛乳カステラ, DAYSAFTER, ローゼンメイデン水銀燈プーリップ, 間桐桜, 牧場物語キミ, ギルティギアイグゼクスアクセントコア, 天羽雅音, 一騎当千呂, FF タクティクス獅子戦争, 諸君私, 来週放送, DEARS 朗読物語, COMEACROSS, ASOSBRIGADE, DL 販売, 人気シミュレーション, 魅力満載, 登校 VER
アフィリエイト系 軽量ボディ, EXILIM, クレバリー, SPEEDUSB, GBHDD 内蔵, PRIUSAIR, ハイビジョン HDD, 特価 COM, ロジテック LHD, GB メモリ搭載, 万画素 CMOS, インチ光沢, PANASONICLUMIX, MYMIODCP, ワイヤレス LAN, DVD コンボ, 薄型デジタル複合機 MYMIO, ブラビアエンジン, 液晶パネル搭載, GB 内蔵, 対応ハードディスク, GIGABEATV シリーズ, ワイド TFT 液晶モニタ, インチワイド TFT 液晶モニタ
お小遣い系 販売マージン, 連複競馬予想, インターネット・コピーライティング基礎講座, アフィリエイト・, 売買指示プログラム, 一流トレーダー, 脱出マニュアル, 法マニュアル, 情報起業成功パーフェクトガイド GOLD, 英語習得, ネット広告ツール, トップセールスマン, 携帯アフィリエイトノウハウ, ナンバーズ攻略, 方法私, 携帯オークション, 勝チ組投資

各コミュニティに対し，より深くコミュニティを分析するために，コミュニティに所属するキーワードのみで再度クラスタリングを行った．この時，「男性コミュニティ」，「オタク系」の２つのコミュニティにおいて，明確な特徴が分かるクラスタリングが行われた[図 3][図 4].

図 3. 男性コミュニティ
（括弧内はキーワード数）

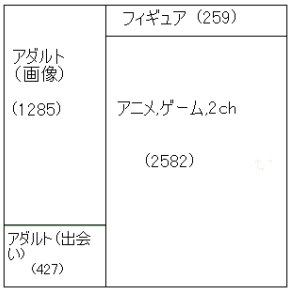

図 4. オタク系コミュニテ
（括弧内はキーワード数）

4.3. オピニオンリーダー抽出

「政治，事件」「アニメ，ゲーム，2ch」「女性」に対し，オピニオンリーダーを抽出した[表 5].

「政治，事件」コミュニティで得られた「zara's voice recorder」は，「教育問題，放送捏造問題，政治問題，耐震偽造，新聞の記事盗用問題」などの問題を「yahoo.co.jp」「asahi.com」「ja.wikipedia.org」などをソースとして，ソースへのリンクと記事の引用を行い，まとめていた．

「アニメ，ゲーム，2ch」コミュニティで得られた「だいちゃんの時間を見つけて書く日記【慢性的寝不足仕事人】」では，「一般，IT，アニメ，ゲーム，漫画， 小説，CD，ラジオ， 映画，ドラマ，イベント，はてな，キーワード，戯言」にカテゴリを分け，「animate.tv」などをソースとして，ソースへのリンクと数行の管理者を記述し，まとめていた．

「女性」コミュニティで得られた「Ｍｙジャニ日記★パート②」では，ジャニーズ関係のタレントの活動を詳細に記述をしていた．ソースへのリンクはなく，自らのファン活動自体がソースであった．

どのブログにも共通していえることは，短い更新間隔と，個人が提供するのには多大な情報量である．生活主体がこれらのコミュニティに立脚し，ブロガーが時間を割き情報量の高い記事を作成していると推測できる．抽出されたおオピニオンリーダーのブログを参照するだけで，各コミュニティの話題を幅広く取得できることができることを確認した．

表 5. 各コミュニティのオピニオンリーダー

政治，事件	zara's voice recorder http://zara1.seesaa.net/ 教育問題，放送捏造問題，政治問題，耐震偽造， 新聞の記事盗用問題などを扱っていた．
アニメ，ゲーム，2ch	だいちゃんの時間を見つけて書く日記【慢性的寝不足仕人】 http://d.hatena.ne.jp/daichan330/ アニメ，ゲームを中心に 様々なニュースを扱っていた．
女性	Ｍｙジャニ日記★パート② http://suk2.tok2.com/user/pk49/ ジャニーズ系の情報を主に 取り扱っている． プロフィールによると， 著者は「15 歳，A 型，女性」.

5. おわりに

本稿では，従来の Web コミュニティ抽出法では解決できなかったブログコミュニティ抽出を，キーワードを分離することにより解決した．キーワードは新鮮さとコミュニティ性を高めるために，「ニュース性のあるキーワード」を抽出した．

提案手法について Newman 法によるクラスタリング実験を行ったところ，7 個のコミュニティが Q=0.5331 という値で得られた．得られたコミュニティの特徴としては，主にパーソナリティの類似性によって結びついており，コミュニティの粒が大きかった．これは「ニュース性のあるキーワード」が普及する範囲が広いために，最大限まで粒が大きくなったと考えられる．女性コミュニティと男性的なコミュニティ 2 つを比べると，女性コミュニティはそれ以上クラスタリングにより分離するのが難しかったのに比べ，男性的なコミュニティはクラスタリングを行うと，興味の対象を得ることができた．このことから，女性は話題をまんべんなく，男性は特定の興味の範囲内だけをブログに記述しているという傾向が推測できる．

また，得られた「オピニオンリーダー」のブログは，どれも積極的に興味の対象の話題を網羅しようとする性質のものであった．

5.1. 課題

実験結果では，コミュニティの粒が大きく，得られたオピニオンリーダーですらもコミュニティ全体の話題を網羅できているとはいえなかった.特に「女性」というコミュニティで得られたオピニオンリーダーは，プロフィールによれば，「１５歳，A 型，女性」であったが，全てのブログに参加している女性がこの著者と類似したプロフィールを持つわけではない．このことから，提案手法でのコミュニティの粒は大きすぎであり，コミュニティの粒を小さく抽出することが課題である．オピニオンリーダーと判定されたブログが，コミュニティ内の話題を網羅できる程度が理想的である．特に Newman 法でクラスタリングを行う場合は(1)Q 値をより高く, (2)コミュニティの数をより多く得られるように，手法を改良していくことが必要である.

次に，本稿の提案手法において，キーワードは名詞列の組み合わせによって求めたが，キーワードは未知語である可能性があるため名詞と判定されない場合がある．そのため，形態素解析に頼らないキーワード抽出手法の開発が課題である.

また，今回の実験は 1 ヶ月間分のデータという比較的短期間だったことから，長期間のデータへの適用も課題である.

謝辞

本研究は株式会社アクセラテクノロジーの萩原様に,多大なるご協力をいただきました.この場を借りて,お礼申し上げます．また,本研究の一部は文科省科学振興費「リーディングプロジェクト e-Society 基盤ソフトウェアの総合開発」によるものである.

文　献

[1] 総務省，“ブログ・SNS（ソーシャルネットワーキングサイト）の現状分析及び将来予測,” http://www.soumu.go.jp/s-news/2005/050517_3.html , 2005.

[2] 奥村学，南野朋之，藤木稔明，鈴木泰裕，“blog ページの自動収集と監視に基づくテキストマイニング,”第 6 回セマンティックウェブとオントロジー研究会，SIG-SW&ONT-A401-01，2004.

[3] Kleinberg, J.M, “Bursty and Hierarchical Structure in Streams,” Proceedings of ACM SIG-KDD2002. pp. 1-25, 2002.

[4] kizasi.jp, http://kizasi.jp

[5] Technorati (テクノラティ) ブログ検索, http://www.technorati.jp/

[6] Nakajima, S.,Tatemura, J.,Hino, Y., and Tanaka, K., “Discovering Important Bloggers Based on Analyzing Blog Threads,” Proceedings of the WWW2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

[7] Nakajima, S., Tatemura, J., Hara, Y., Tanaka, K., and Uemura, S.,“Identifying Agitators as Important Blogger based on Analyzing Blog Threads,” Lecture Notes in Computer Science 3841, The 8th Asia-Pacific Web Conference (APWeb2006), pp. 285-296, 2006.

[8] Kumar, R. and Raghavan, P. and Rajagopalan, S. and Tomkins, A., “Trawling the Web for emerging cyber-communities, ” Proceeding of the 8th International Conference on World Wide Web table of contents, pp.1481-1493, 1999.

[9] Kleinberg, J.M. and Kumar, R. and Raghavan, P. and Rajagopalan, S. and Tomkins, A., A. , “The Web as Graph: Measurements, Models, and Methods, ” COCOON’99, LNCS 1627, pp.1-17, 1999.

[10] 村田剛志，“参照の共起性に基づく Web コミュニティの発見,” 人工知能学会論文誌, Vol.16,No.3, pp.316-323, 2001.

[11] 谷口智哉，松尾豊，石塚満，“Blog コミュニティの抽出と分析,” 人工知能学会: 第 6 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A401-08, 2004.

[12] Newman, MEJ and Girvan, M.,“Community structure in social and biological networks,” Proceedings of the National Academy of Sciences, Vol. 99, No.12, pp. 7821-7826, 2002.

[13] 内田 誠，柴田 尚樹，“ブログ記事ネットワークからの emering topic の抽出と可視化,”人工知能学会第 20 回全国大会, 2006.

[14] MeCab, http://mecab.sourceforge.net/

[15] Newman, MEJ and Girvan, M.,“Finding and evaluating community structure in networks,” Physical Review E. 69, pp.26113-26128, 2004.

[16] Newman, MEJ,“Fast algorithm for detecting community structure in networks,” Physical Review E. 69, pp.066133-066138, 2004.

[17] Google BlogSearch, http://www.google.co.jp/intl/ja/help/about_blogsearch.html